Intelligent Embedded Vision for Summarization of Multi-View Videos in IIoT

Tanveer Hussain, *Student Member*, IEEE, Khan Muhammad, *Member, IEEE*, Javier Del Ser, *Senior Member, IEEE*, Sung Wook Baik, *Member, IEEE*, Victor Hugo C. de Albuquerque, *Member, IEEE*

***Abstract*—Nowadays, video sensors are used on a large scale for various applications including security monitoring and smart transportation. However, the limited communication bandwidth and storage constraints make it challenging to process such heterogeneous nature of Big Data in real time. Multi-view video summarization (MVS) enables us to suppress redundant data in distributed video sensors settings. The existing MVS approaches process video data in offline manner by transmitting it to the local or cloud server for analysis, which requires extra streaming to conduct summarization, huge bandwidth, and are not applicable for integration with industrial internet of things (IIoT). This paper presents a light-weight CNN and IIoT based computationally intelligent (CI) MVS framework. Our method uses an IIoT network containing smart devices, Raspberry Pi (clients and master) with embedded cameras to capture multi-view video (MVV) data. Each client Raspberry Pi (RPi) detects target in frames via light-weight CNN model, analyzes these targets for traffic and crowd density, and searches for suspicious objects to generate alert in the IIoT network. The frames of each client RPi are encoded and transmitted with approximately 17.02% smaller size of each frame to master RPi for final MVS. Empirical analysis shows that our proposed framework can be used in industrial environments for various applications such as security and smart transportation and can be proved beneficial for saving resources[1].**

***Index Terms*—Artificial Intelligence, Big Data, Convolutional Neural Network, Computational Intelligence, Computer Vision, IIoT, IoT, Video Summarization.**

I. Introduction

The striking advancements in communication and networking technologies since the birth of the Internet yielded several forms of network architectures and gigantic size of data. A fertile product of this evaluation, Internet of Things (IoT) [1-4], is developing very rapidly and has replaced the traditional sensing of surrounding environments. The IoT is a smart wireless network [5], integrating many agile devices for exchange of information, communication with each other, and generates huge amount of data on a daily basis. The video data generated by vision sensors in IIoT, installed in industries particularly meet the requirements of Big Data [6] and are becoming the key sensor devices for various IIoT applications. A single vision sensor in an IIoT with 25 frames per second (fps) generates huge amount of video data hourly. As IIoT is the integration of interconnected smart devices, which shows that there are multiple cameras installed at different places and yields big industrial video data. The amount of cameras exponentially lifts up the magnitude of video data, making it Big Data repositories. This data requires efficient processing in industries for several purposes such as employee's monitoring and salient events detection. The basic requirements of such data in industries include redundancy removal along with presentation and preservation of only salient data in compact form for future use and analysis.

The video data generated from IIoT [7] are used for various applications [8, 9] such as security monitoring, tracking [10], and intelligent transportation. The mainstream devices connected in IIoT are resource constrained [11] with limited computation power and storage which cannot process such big video data. Thus, the generated data are possibly transmitted to cloud with unlimited computational and storage resources for further analysis. Cloud computing is considered as suitable place to analyze such Big Data efficiently [12]. However, the IIoT systems are expanding leaps and bounds, generating huge amount of video data which requires quick response for public safety [13], healthcare [14], smart industries [7, 15], intelligent transportation [16, 17], and emergency situations handling. The issue with cloud-based solutions is that they are offline and lack reliability. There is always a huge IIoT traffic over wireless networks that are well known with properties such as low bandwidth and high communication cost [18]. It is obvious from the facts that efficient, real-time, and reliable decision from IIoT vision sensors data is not guaranteed via cloud computing. Furthermore, along with processing such Big Data, storing it for future use is also a big challenge. It is very difficult in an IIoT environment for a resource constrained device to store 90,000 frames for a single camera and 360,000 for a network of four cameras per hour. It requires huge storage devices, however, that is not feasible in an IIoT environment. Thus, keeping only important and representative information of whole day lengthy videos is a better option in terms of limited storage in IIoT.

Video summarization is an automatic technique for extracting significant information from big video data in the form of keyframes or video skims [19]. It investigates the input video for different events, informative frames, and generates a summary that is representative of the whole video. Video summarization is broadly divided into two categories: single-view video summarization (SVS) and MVS. SVS generates

Manuscript received May 10, 2019; Accepted: XXX, Published: XXXX. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2019R1A2B5B01070067). This paper was recommended by Associate Editor XYZ. (*Corresponding author: Sung Wook Baik*)

Tanveer Hussain, and Sung Wook Baik are with Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, South Korea (Email: tanveer445@ieee.org and sbaik@sejong.ac.kr)

Khan Muhammad is with the Department of Software, Sejong University, Seoul 143-747, South Korea. (Email: khan.muhammad@ieee.org)

Javier Del Ser is with TECNALIA, 48160 Derio, Spain, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain, and Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Spain (Email: javier.delser@tecnalia.com)

Victor Hugo C. de Albuquerque is with Graduate Program in Applied Informatics at the Universidade de Fortaleza, Fortaleza/CE, Brazil (Email: victor.albuquerque@unifor.br)

[1] https://github.com/tanveer-hussain/Embedded-Vision-for-MVS

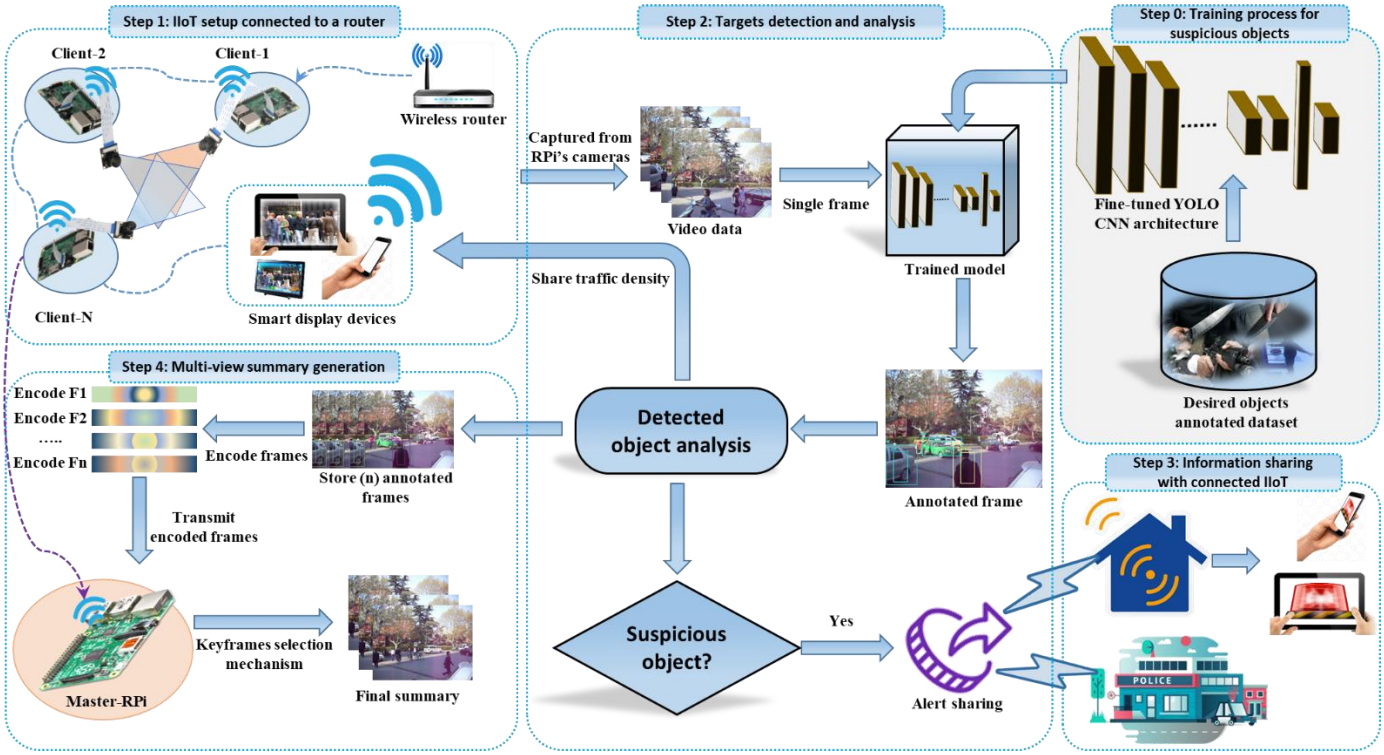


Figure 1: Overall proposed framework with different steps for training, data acquisition, objects analysis, and summary generation in an IIoT. **Step 0**: This is an offline step which fine-tunes an existing object detection model for the desired objects (i) persons, (ii) vehicles, and (iii) suspicious objects (guns, knives etc.). **Step 1**: IIoT setup with client RPi's and smart display devices connected to a wireless network in an industry. The embedded RPi cameras generate video data. **Step 2**: It receives the video data, passes a single frame to the trained model which outputs an annotated frame whose objects are analyzed for (i) traffic density to share with connected IIoT devices/administration (ii) suspicious objects, and if there is any, an alert is generated. **Step 3**: The alert received from step 2 is shared with the concerned departments and smart industries in IIoT environment. **Step 4**: It receives (n) number of annotated frames that contain dense targets which are encoded and transmitted to the master RPi in the same network for keyframes selection.

summary from a single input video, covering only one view at a time. A single camera for smart industries in IIoT network has a limited coverage that cannot fully exploit the overall environment synchronously. MVS considers different view-points in parallel and generates representative summary of all the views. As opposed to SVS, MVS is challenging because it has both inter- and intra-view correlations to be considered while generating summary. Another big challenge in MVS is variation of light conditions among different views. Furthermore, synchronization among different views also makes the problem of MVS a difficult one. The research domain of MVS has wide range of applications including both indoor and outdoor environments. It can be used for automated monitoring in smart industries, human action and activity recognition, intelligent transportation, security and law enforcement [20], sports, and entertainment [21]. The basic flow of MVS comprises of three steps: preprocessing MVV, features extraction, and summary generation. The first step in MVS flow suppresses the MVV through several redundancy removal techniques such as shots segmentation (uniform or variable length) and video splitting based on the shots boundary. Features extraction, objects detection, and tracking are considered to be the second prerequisite step followed by different MVS methods. Final step of MVS involves summary generation from the extracted features through various machine learning or template matching algorithms.

The literature of MVS follows traditional machine learning, neural networks, motion, and saliency based techniques. Majority of the MVS methods use low-level features and clustering techniques to generate summary. The first research of MVS [22] represented correlations among different views through a hyper-graph. They converted the hyper-graph into a Spatio-temporal graph and finally used random walks clustering on it to generate summary. Li et al. [23], inspired from [24], used maximal marginal relevance for MVS. which compensates the relevant keyframes and ignores the redundant ones. The idea proposed in [23] used a bandwidth-efficient and online keyframes extraction method. They utilized K-means clustering for final summary generation. Mahapatra et al. [20] presented human actions recognition based technique for MVV synopsis. A deep learning based method where learned features are extracted from frames and preserved in embedding space for intra-view correlation computation is presented in [25]. They used "BVLC CaffeNet" model which is trained on large-scale ImageNet dataset with millions of images for deep features learning. The frames with similar features are kept close to each other and summary is generated using latent space clustering. Panda et al. [26] introduced C3D [27] features in MVS domain for shots segmentation. In this method, they computed two proximity matrices to seek intra- and inter-view correlations. Next, they calculated the pairwise Euclidean distance between frames and finally generated summary through sparse representation selection over learned embedding. The most recent method proposed by Hussain et al. [28] makes use of complex CNN for sequential features

extraction followed by deep bi-directional LSTM for final skims generation.

The analysis of employed MVS techniques covered in aforementioned paragraph reveal the limitations of these techniques and are discussed here one by one. First of all, mainstream MVS techniques use traditional low-level features for shots segmentation which result in a non-representative summary of the input MVV. Secondly, all the employed techniques for MVS are not online, thus, waste bandwidth and time of transmitting video data to cloud or local computers of the connected network to generate summary. The existing MVS techniques have no special mechanism for detecting suspicious objects before the summary generation to take preventive actions instantly. Another key limitation of all employed MVS approaches is their offline processing that leads to a slow response to any abnormal action or activity. Similarly, all the discussed methods are only functional with a lot of hardware requirements i.e., cameras and wireless sensor networks, computer systems, and cloud storages in some cases. Furthermore, the existing techniques do not focus on the salient objects such as human, vehicles, and suspicious objects that help in further analysis of the video data for various applications.

To tackle all these issues, we introduce an intelligent IIoT based framework with embedded vision for suspicious objects detection [29], traffic density information sharing, and MVS. The key contributions of our system are summarized as follows:

1. MVS processes huge amount of video data generated from distributed video sensors. Majority of existing MVS techniques receive multi-view videos through wireless network and generate offline summary. This requires maximum communication bandwidth and wastes storage capacity. In this paper, we present an MVS technique which can generate online summary, compress the keyframes, and transmit them to smart devices connected in IIoT network for further analysis. Thus, our system saves communication bandwidth and provide online MVS.
2. It is very hard to install cameras, connect them with computers through wires, and monitor videos manually for suspicious objects in industries. Further, the majority of available MVS techniques are very expensive in terms of processing time and hardware implementation which rely only on summary generation. In this framework, we tackle the problem of hardware implementation by installing an embedded device with camera and for suspicious object detection problem we investigate light-weight CNNs for efficiency. The key contribution of our system is installing a single hardware device capable of detecting and reporting about suspicious objects and traffic density to authorities in industrial setup.
3. The huge amount of video data generated by distributed video sensors need CI algorithms for efficient processing. To achieve the goal of dealing Big Data with CI algorithms in IIoT precisely, we employ efficient and light-weight CNN to suppress the redundant video data. Our proposed framework reads input video (6-fps) and discards the frames with no salient objects. The salient objects in industrial surveillance Big Data are vehicles, persons, and suspicious objects such as knife, gun etc. Thus, we deal with Big Data in IIoT through CI techniques for accurate reduction of data which assist in further steps and decreases time complexity.
4. The currently employed MVS techniques are intractable to be integrated with IIoT and many other smart devices. Our proposed framework can be integrated with IIoT and the output can be easily observed via any smart device. If a device is connected to the said wireless network, it can be accessed anytime and anywhere without sitting in a special surveillance monitoring room. So, we contribute to MVS literature by presenting a framework that is adaptable and can be used in any IoT/IIoT environment for summary generation.

The remainder of this paper is organized as follows. The overall proposed framework is explained in Section II. In Section III, the experimental results for Road/Office dataset and YouTube videos along with running time of embedded device is given. In Section IV, this research work is concluded with some limitations of current framework and future plans for a better MVS method.

II. Embedded Vision based MVS Framework in IIoT

In this section, the working procedure of our proposed MVS framework in IIoT environment for smart industries is explained in detail. There are four main steps that are performed online and one training step which is executed offline to acquire a trained model. In the offline step, we fine-tune an existing model for three type of objects detection. In the online steps, first we have an IIoT setup having client RPi's with attached cameras and smart devices connected to the same wireless network. Each client RPi captures video data, inputs it to CI neural network based object detection model to attain annotated frame. The client RPi analyzes the annotated frame for suspicious objects and traffic density and stores (n) number of frames in its memory. Finally, the stored frames are encoded and transmitted to the master RPi in the same network to generate summary. All these steps are visualized in **Fig. 1** and explained in detail in coming sections.

A. Training Object Detection Model

Recently, CNNs showed an outstanding performance for various tasks such as classification, segmentation, retrieval, and object detection. It has been widely used for many applications [30] including action and activity recognition [31, 32], security [33] and many others. Therefore, motivated from these studies we investigated CNNs for our problem in IIoT environment to detect suspicious objects for instant reporting. The computational complexity of CNNs is a big hurdle to practice CNN based intelligent algorithms over resource constrained devices. To tackle this challenge, we chose a precise, light-weight, and efficient CNN model for object detection. We fine-tune an existing object detection CNN model to detect only three type of objects that can be utilized for summary generation and further analysis. The three target objects are vehicles, suspicious objects, and persons. Vehicles include all types of buses, cars, and bikes etc. Suspicious objects used for training are guns [29], pistols, knives, and almost all possible weapons that can be alarming in industries. We used YOLOv3 tiny [34] object detection model in our proposed framework. It is more

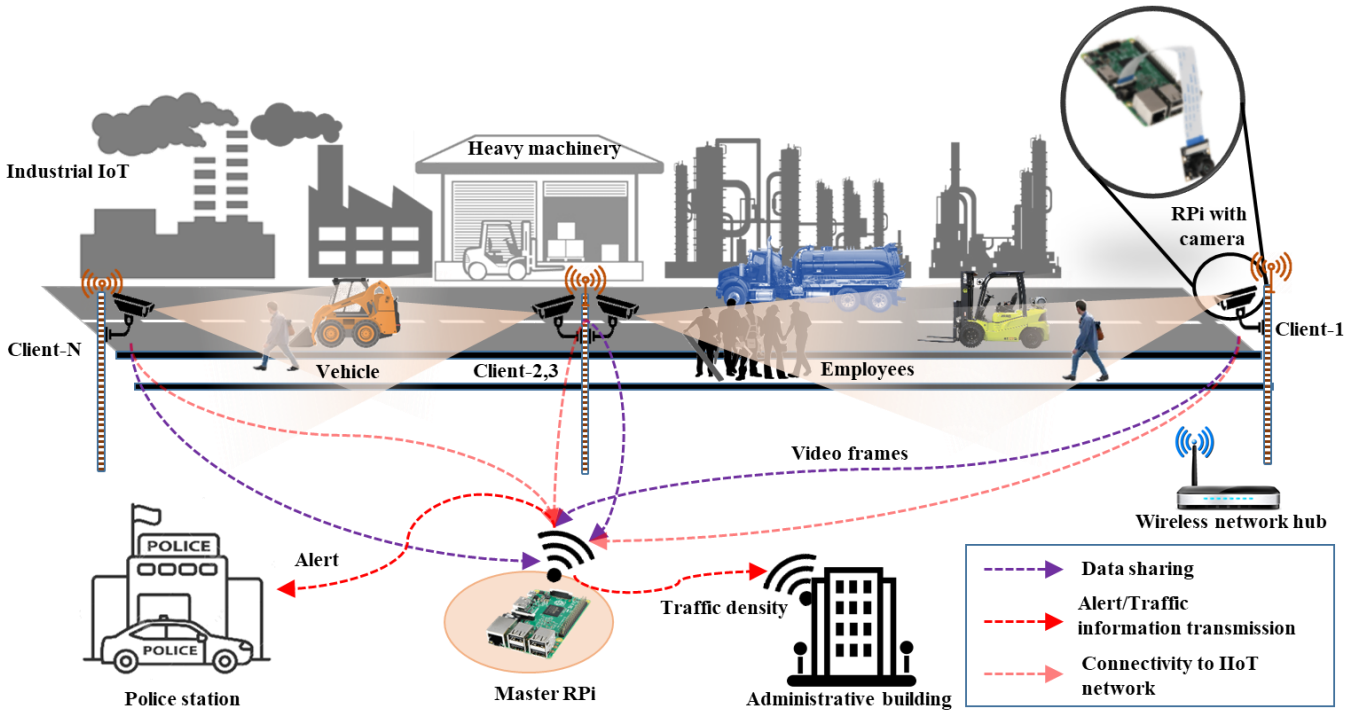


Figure 2: Sample scenario for IIoT connected devices (RPi's) in smart industries. The client RPi's are focusing on different views, intelligently monitoring the environment for suspicious objects, and ready to send alert signal to the concerned department. The client RPi's also share information about traffic continuously with the users in smart industry's administration.

than 100 times faster compared to a famous objected detection CNN model Faster-RCNN [35]. As we used tiny version of YOLO which can process input frames on RPi efficiently. Therefore, we converted the datasets to 6 fps for the input videos so that RPi can process it easily. In the training process, we first acquired data for each class that is useful in monitoring of smart industries and its details are given in experimental section. Next, we annotated the data manually by selecting the area of the object and stored it in a text file having number of objects along with their locations inside image. The text file is then converted into YOLO data format. Finally, we inputted the images, modified configuration file, and pre-trained weights to the training function of YOLO that stores the updated model after every 100 iterations. We modified the configuration file of tiny YOLO with our desired number of classes and changed the count of filters according to our data. As there are three type of objects, so we changed the filters to 24 by formula (classes + 5) ×3.

B. *IIoT Setup Connected to a Router*

The future part of Internet is IoT including millions of connected sensors such as vision, GPS, and traffic sensors etc. for communication, computation, and intelligent monitoring of smart industries. Normally in industries video data from vision sensors is transmitted to cloud for better understanding and efficient analysis, requiring uploading data to cloud. The data transmission to cloud yield huge wastage of communication bandwidth, time, and makes real-time response impossible corresponding to abnormal actions or activities. To overcome these challenges, we propose a novel resource constrained RPi based framework. Vision sensor is attached with each RPi and is connected to a wireless network in IIoT environment for efficient and intelligent processing of video data. The overall scenario of the proposed framework in IIoT setup is presented in **Fig. 2.**

The client RPi's can be installed at locations where multi-view data are important to be captured in smart industries. There can be a network of (n) number of RPi's inter-connected to capture video data individually. Each client RPi captures video data at 6-fps, to make the process online and efficient. The client RPi then passes a single frame to the object detection model that annotates it for the three desired objects. Detected objects are then analyzed for targets detection explained in coming section.

C. *Targets Detection and Analysis*

The input frames acquired from sub-section (II.B) are processed in this step. The trained model in the initial step of our framework is used to annotate frame for three kind of objects: vehicles, persons, and suspicious objects. The density of vehicles and persons is computed from the annotated frame and shared with the smart devices connected to the IIoT network. Information about traffic in smart industries helps building up a routine and plays a key role in saving time. If the annotated frame contains any type of suspicious object such as gun or knife it is considered as an alert. The alert is shared with the concerned devices in IIoT for employee's safety and also with the police department for quick preventive actions. The police department can analyze the situation of alert by watching the camera and if it is not alarming, they can ignore it. After a continuous detection of suspicious object, an alert is sent again to the police department.

D. *Multi-view Summary Generation*

The final summary is generated on master RPi. The input to this step is (n) number of frames from each client RPi. We select n to be 10. Therefore, we encode 10 frames from each view having many targets like vehicles and persons. We apply lossless PNG compression on these frames and send it to the master RPi. The PNG compression has the advantage of saving communication bandwidth. The master RPi receives PNG compressed frames in the form of a vector and decodes it to restore the original frames. The same process is applied for rest of the frames and after decoding they are processed via two methods: entropy and complexity to compute the information present in frames. Procedure of each method is explained in detail in subsequent sub-sections. These methods output a real-value between 0 and 1, and the frames with highest value of sum for complexity and entropy is selected as a keyframe. The number of keyframes in single chunk of 10 frames depends upon the user threshold. Currently we select single frame per-view from the frames received in a single chunk. The total frames in a single chunk are (n * number of client RPi's), where n is the number of frames each client RPi's transmits. **Fig. 3** represents the values of entropy and complexity corresponding to the frames having high and low information with and without compression. The results are same for the selection of keyframes indicating no effect of the compression scheme applied over the frames.

Frame #	Entropy value	Complexity value	Sum	View
150	0.2997	0.5100	*0.8097*	2
198	0.2986	0.5002	0.7988	2
1224	0.3013	0.5476	**0.8489**	3
1236	0.2811	0.5044	0.7855	1
	Decoded frames (After applying PNG compression)			
150	0.2987	0.5002	*0.7989*	2
198	0.2975	0.4970	0.7945	2
1224	0.2993	0.5459	**0.8452**	3
1236	0.2833	0.5035	0.7868	1

Figure 3: Sample video frames from Road dataset videos of different views. **Frame #150:** higher entropy and complexity values show high amount of information present in frame and the objects detected are near to the camera which means they are important and need to be considered for the final summary. **Frame #198**: the complexity and entropy values are lower because of low amount of information and objects are far from the camera. **Frame #1224**: The objects are nearer to the camera and frame contains higher information. **Frame #1236:** Although there are many vehicles in this frame but they are far from camera so the information noticed is comparatively lower than **Frame #1224**. The keyframes before and after applying encoding/compression scheme are the same. The sum of entropy and complexity for different frames is given in the table where for both the scenarios (before and after compression and decoding) the two keyframes among the four are the same (**Frame #150 and Frame #1224**). Frames with highest sum are made bold and the second highest sum is made italic and underlined inside the table in figure.

1) *Entropy*

Image entropy indicates the amount of information inside a frame. The higher value of entropy represents that the frame is rich of information and lower-value of entropy shows that the frame has less amount of information. The process flow of entropy value computation is given in **Fig. 4 (a)**.

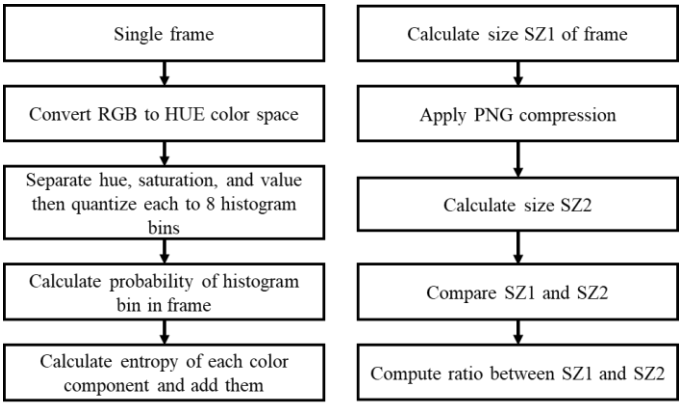


Figure 4: **(a)** Computing entropy score and **(b)** Computing complexity score from a single input frame.

2) *Complexity*

The complexity of a frame is computed by comparing original size of the frame and its size after PNG compression [36] is applied. The PNG compression algorithm is applied according to the human visual system as explained in [37]. High compression rate shows that the frame has little visually important information and vice versa. Complexity value computation is visualized in **Fig. 4 (b)**.

III. Experimental Results

In this section, we evaluated our proposed framework using empirical analysis, subjective, and objective evaluation. We evaluated the performance of client and master RPi in IIoT environment in terms of saving communication bandwidth and transmission time through empirical analysis. The subjective evaluation is performed through a user's case study and objective evaluation is done over an existing MVS dataset. Evaluation of the proposed framework is explained in detail in next sub-sections.

A. *Experimental Settings*

The experimental setting contains several client RPi's for video data acquisition and a single master RPi for summary

generation connected in an IIoT network. The whole framework is implemented in Python (version 3.5) with optimized OpenCV (version 3.4.0) build for our specific configuration. The RPi used is model B (version 3) having ARM Cortex A53 processor type with 1.2 GHz speed. The RPi is equipped with 1 GB main memory. Further explanation about hardware architecture of used RPi is out of scope of this paper.

B. Data Acquisition

The data used for training and experiments is acquired from different sources. The primary data used for training are downloaded from YouTube and captured screenshots from some movies. The MVS of the proposed framework is evaluated using Road [22] and Office [22] benchmark datasets that are publicly available.

1) Road Dataset

Road dataset is the most challenging one in MVS literature. It has no synchronization among different views. The camera used to capture video data is handheld with high-level of shuddering. The videos of this dataset also suffer from light variations among different views. This publicly available dataset has no ground truth available. Therefore, we evaluated our experiments using subjective evaluation. As our framework is also feasible in outdoor scenarios with vehicles, persons, and suspicious objects, accordingly we utilized road dataset for our experimental evaluation. It is recorded in an outdoor environment with persons and vehicles moving on the road. Thus, along with MVS we used this dataset for evaluation of traffic density module of our proposed framework.

2) Movies and YouTube Data

The fine-tuning of an existing object detection model YOLOv3 tiny needs extensive data for accurate training. We found no particular benchmark dataset for suspicious objects in surveillance. Therefore, we collected our own data for the training of object detection model. The data collected form YouTube lack some sort of weapons that are important to be considered as suspicious. Thus, we captured screenshots from some action movies having these sort of weapons to make our dataset complete.

C. Communication Bandwidth and Transmission Time Analysis

The main objective of our framework is to save as much communication bandwidth and storage in IIoT environment as possible. The deep power analysis of saving bandwidth is out of scope of this paper and here we only provide the technical details about saving communication bandwidth and overall transmission time. Since we proposed framework for IIoT environment with different devices connected together via wireless networks. Thus, it is very important to analyze communication bandwidth and transmission time. The traffic over such networks is dense and the wireless networks have weak property of limited bandwidth and communication cost. Thus, efficient communication is a basic requirement of a wireless network and our proof for claim of saving bandwidth is shown in **Fig. 5.**

To prove the concept of saving communication bandwidth and transmission time we make an analysis for the frame in **Fig. 5 (a).** It is clear that a file of 1.99 MB size consumes 1.99 MB communication bandwidth which is far greater than 0.34 MB that is transmitted by our proposed framework. We transmit the frame in encoded format from client to master RPi. Let's suppose we have ideal situation in the network and we want to transmit a frame from client to master RPi where the distance between client and master is 0.3 Km and assume that the speed of the signal is 200,000 Km/s with data rate of 32 Mbps. In order to reach the destination (master RPi) from source (client RPi) there are three types of transmission times involved. First is to get the frame on the ne$_t$work from RPi, second is to transmit it through distance of 0.3 Km to destination, and third is to put the data into the master RPi from the network.

$$t_1 = \frac{data\ size}{data\ rate} \tag{1}$$

$$t_2 = \frac{distance(client\ to\ master\ RPi)}{transmission\ speed} \tag{2}$$

Total time taken to reach from client to the master RPi is the sum of these three individual transmission times. The time taken from source to the network is given in Eq. 1 and transmission time on the network is calculated via Eq. 2. The third time is the same as the first transmission time. A detailed analysis and calculation is provided in **Table I.**

TABLE I
DETAILED DESCRIPTION OF TIME ANALYSIS FOR A SINGLE FRAME TRANSMISSION FROM FIG. 6 (A). IT CAN BE OBSERVED FROM THE TABLE THAT TIME DELAY IS 0.8678539. THE DIFFERENCE OF SIZE BETWEEN THE TWO (ORIGINAL AND ENCODED) FRAMES IS APPROXIMATELY 17.02%.

Frame nature	**Size of data (MB)**	**Transmission time (secs)**	**Remarks**
Encoded/ Compressed	0.3415	0.0895221	From client RPi to network
		0.0000015	Network processing time
		0.0895221	From network to master RPi
	Total time	**0.1790458 secs**	
Original	1.9968	0.523449139	From client RPi to network
		0.0000015	Network processing time
		0.523449139	From network to master RPi
	Total time	**1.0468997 secs**	

Description	Resolution	Dimension	Size (MB)	Remarks
Original image	1040 x 1920	3-D	1.9968	Compression ratio is 0.17 indicating less information
Encoded image	341591 x 1	1-D	**0.3415**	
Decoded image	1040 x 1920	3-D	1.9968	

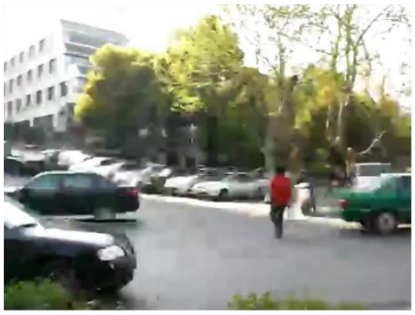
(a)

Description	Resolution	Dimension	Size (MB)	Remarks
Original image	720 x 1280	3-D	0.9216	Compression ratio is 0.51 indicating high information
Encoded image	477909 x 1	1-D	**0.4779**	
Decoded image	720 x 1280	3-D	0.9216	

(b)

(c)

(d)



Description	Resolution	Dimension	Size (MB)	Remarks
Original image	480 x 640	3-D	0.3072	Compression ratio is 0.25 indicating normal information
Encoded image	78153 x 1	1-D	**0.0781**	
Decoded image	480 x 640	3-D	0.3072	

Description	Resolution	Dimension	Size (MB)	Remarks
Original image	480 x 640	3-D	0.3072	Compression ratio is 0.19 indicating less amount of information
Encoded image	58858 x 1	1-D	**0.0585**	
Decoded image	480 x 640	3-D	1.9968	

Figure 5: Sample results for claim of saving communication bandwidth and transmission time in an IIoT network. (a) a frame from a movie with a person lifting gun is compressed from 1.99 MB to 0.34 MB for transmitting it through wireless network in IIoT, (b) the frame from YouTube video is compressed from 0.921 MBs to 0.477 which can be helpful in decreasing the network traffic, as IIoT network contains a lot of exchange of information, (c) a frame of Road dataset is compressed from 1.99 MB to 0.07, and similarly (d) represents a single frame from Road dataset video with 0.19 compression rate indicating how useful it is to compress a frame and send it through wireless network and then decode it. The smallest size in MBs in the Tables is bold and underlined.

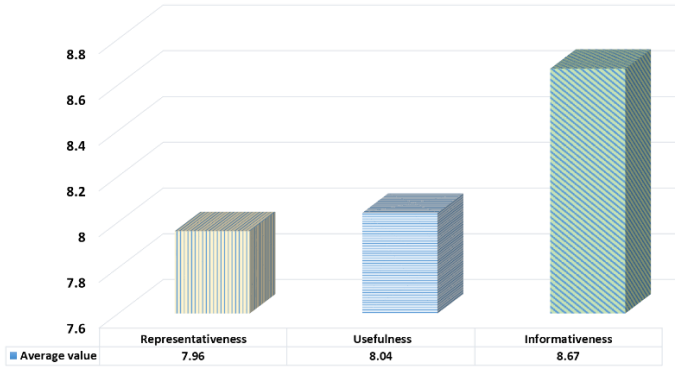


Figure 6: Average values of all the participants given score. It can be observed that our proposed method achieves better results for Informativeness and Usefulness which are very important for any MVS method to preserve these properties. The value of Representativeness is although comparatively low but still is reasonable.

D. Subjective Evaluation

To evaluate the effectiveness of our proposed framework, we made a user survey. In this survey we invited 12 Master and PhD students. The students we selected for participation are well known to the computer field and especially computer vision. The age of these students ranges from 20 to 26 years. The basic aim of this survey is to check whether our proposed framework is able to extract the most informative frames of multi-view videos or not. Furthermore, we want to confirm that our method is able to compute inter-view correlations intelligently without involving any extra processing complexity. We provided the generated summary of our proposed framework, videos of road dataset with 6 fps, and a spreadsheet with descriptive and multiple-choice questions to the participants. To avoid any biasness, we gave the same fps videos which we used for experimentation i.e., 6 fps. The road dataset has no ground truth along with the videos, therefore we did not provide any ground truth summary. The overall survey is executed offline with no restriction of time on the participants. The time restrictions make the volunteer pressurized and can do any mistake while performing evaluation. We asked 2 subjective and 3 objective type questions with four choices (strongly agree, agree, disagree, strongly disagree).

TABLE II

DETAILED SURVEY OF 12 PARTICIPANTS. THE SCORE IS ASSIGNED BETWEEN 0 AND 10 WHERE 0 SPECIFIES 'STRONGLY DISAGREE' WHILE 10 INDICATES 'STRONGLY AGREE'. EACH USER IS ASKED THREE QUESTIONS IN WHICH FIRST ONE SHOWS IF THE SUMMARY IS REPRESENTATIVE OR NOT. SECOND QUESTION ASKS ABOUT THE DIVERSITY AND USEFULNESS OF THE GENERATED SUMMARY. THIRD QUESTION IS ABOUT THE KEYFRAMES WHETHER THEY ARE INFORMATIVE OR NOT.

Participant #	**Questions**		
	Representativeness	*Usefulness*	*Informativeness*
1	7.5	8.5	9
2	8	8	9
3	7	8	8
4	8.5	9	8
5	8	8	10
6	9	8	9
7	6	7	7
8	7	6	8
9	9	8	9
10	8.5	8	10
11	8	10	8
12	9	8	9
Average	**7.95**	**8.04**	**8.66**

The questions asked in the survey are inspired from the user case study provided in [22]. Q1: Is there any limitation of the current summary? If yes, write about it. Q2: Is the generated summary enough to represent the input video? The user survey given in **Table II** are objective answers of the participants and the subjective answers of participants are not discussed because of its lower importance. In **Table II**, representativeness refers to the nature of the summary that whether it is able to represent the overall video well enough or it fails to do so. Usefulness indicates that how useful the summary was and whether it can the alternative of the input video or not at all. Informativeness in **Table II** indicates the case where the automatic generated summary contains enough information or the keyframes lacks it. The Q1 in objectives is: Is the summary representative of all the videos? Q2: Would you prefer to keep the summary in your computer instead of the three videos? Q3: Do you think the generated summary contains salient information? The average values of all the participants for three questions with given scores are visualized in **Fig. 6.**

A. Objective Evaluation

To compare the results of our framework with existing techniques we used Office [22] dataset for evaluation. It is a challenging dataset in MVS domain due to variable light conditions among different views and lack of synchronization. It is recorded using 4 stably-held cameras in an office. The precision, recall, and F1 score are used for comparison with the state-of-the-art. Precision captures the ability of any VS technique for removing useless information. The value of recall shows the strength of keeping salient information for any VS method. Detailed comparison with state-of-the-art is given in **Fig. 7**. It is clear from the figure that our method achieves the best result in terms of F1 score with lower computational power requirements. The most recent method [28] uses computationally complex CNN with dense LSTM structure to generate summary. While our method uses resource constrained device with simple and efficient mechanism for MVS. A lag in the value of precision from some of the methods indicates the abundance of extra keyframes.

A. Time complexity

The variation among total execution time and number of processing frames with respect to different fps for each task from the three views of road dataset are given in **Table III**. Decrease in number of processing frames has a positive effect on the total execution time for all the steps i.e. it decreases the time complexity. Further details about the running time of our proposed framework are given in **Table IV**. The specification for both client and master RPi are the same. Client RPi's transmit data to the master RPi for summary generation. The road dataset videos are utilized for time complexity analysis.

TABLE IV

EXECUTION TIME OF RPI'S CORRESPONDING TO THE TASK PERFORMED. OBJECT DETECTION IS ACHIEVED THROUGH LIGHT-WEIGHT CNN OVER CLIENT RPI AND SIMILARLY ENCODING IS DONE OVER THE SAME DEVICE. COMPLEXITY, ENTROPY, AND SUMMARY GENERATION ARE PERFORMED OVER MASTER RPI DEVICE.

Module	**View**	**Video length**	**Task(s)**	**Execution time (secs)**
Client RPi-1	1	12 min, 58 secs	Object detection and analysis, encoding frames	3005.12
Client RPi-2	2	22 min, 02 secs		5112.18
Client RPi-3	3	21 min, 56 secs		5639.49
Master RPi	All	Frames with objects	Complexity value calculation for single frame	0.07959
Master RPi		Frames with objects	Entropy value calculation for single frame	0.32685
Master RPi		Frames with objects	Keyframes selection (summary generation)	247.51

TABLE III

A COMPARATIVE ANALYSIS OF DIFFERENT FPS AND THE RESPONSE OF OUR EMBEDDED DEVICE IN TERMS OF NUMBER OF PROCESSING FRAMES. THE VIDEOS SELECTED FOR EXPERIMENTS ARE FROM ROAD DATASET. THE TOTAL TIME IS FOR COMPRESSION (ENCODING), ENTROPY, AND COMPLEXITY VALUE COMPUTATION. THE BEST EXECUTION TIME IS MADE BOLD.

# View	**Video length (secs)**	**6-fps**		**15-fps**		**30-fps**	
		# processing frames	**Total time (secs)**	**# processing frames**	**Total time (secs)**	**# processing frames**	**Total time (secs)**
1	778	4668	**3781.08**	11670	9452.7	23340	18905.4
2	1322	7932	6424.92	19830	16062.3	39660	32124.6
3	1316	7896	6395.76	19740	15989.4	39480	31978.8

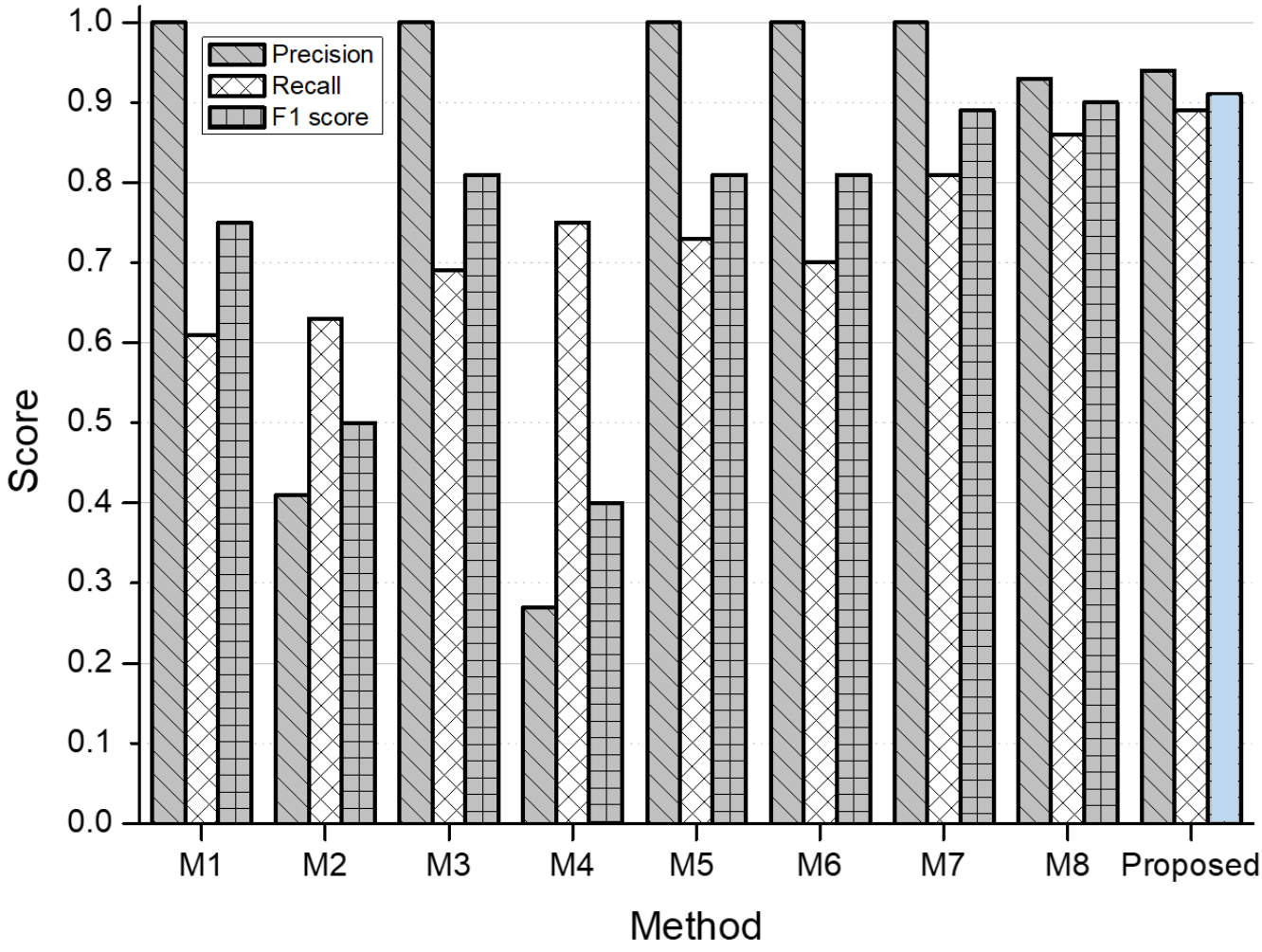


Figure 7: Comparison of our proposed framework with existing MVS methods. The MVS methods are sorted year wise in the figure, M1 [22], M2 [38], M3 [39], M4 [40], M5 [25], M6 [26], M7 [41], and M8 [28]. The proposed method achieves the highest F1 score among existing methods and filled blue.

IV. Conclusions, Limitations, and Future Work

The IoT is advancing very readily and is replacing the traditional sensing of existing systems. IoT is a network of interconnected devices and sensors providing various services. The vision sensors in IoT are emerging recently because of its high-level usage in smart cities/industries for several applications such as security. The video data generated by these sensors is very huge, meeting the requirements of Big Data. The vision sensor data contains redundancy with salient events happening very rare. The redundant data need to be discarded because of computational complexity and the limited storage. Therefore, to solve these issues, we propose a novel CI based online system for video data analysis generated from multi-vision sensors in IIoT in smart industries. We used no special sensors but utilized embedded vision sensor to capture video data and process it intelligently. RPi's with attached cameras are used to capture multi-view data to cover the overall scenes. The video data captured are processed online on each RPi to search for suspicious objects and analyze the traffic situations. The traffic information is shared with desirable connected devices in IIoT and if there is alert about suspicious objects, it is shared with the concerned department for quick actions. Each client RPi transmits the frames with huge density of persons and vehicles to the master RPi for MVS. The frames before transmission over network are encoded through PNG lossless compression to save time, energy, and communication bandwidth. Currently, 10 frames are stored in client RPi and transmitted to master RPi which can be changed according to the requirements. The master RPi that is connected to the same network first decode the compressed frames and then compute information of each frame of all the client RPi's for keyframes selection. The frame with highest information is stored as keyframe and the rest of redundant frames are discarded. Thus, the final summary is generated on master RPi which is very compressed form of the whole video data. Empirical experimental evaluation of our proposed framework substantiate it as an appropriate candidate to be implemented in IIoT scenarios in smart industries for 360-degree coverage and instant response after anomalies detection.

The current system has no appropriate mechanism for anomalous behavior detection of persons. In the future work we want to extend this framework by introducing some mechanisms that can analyze the detected persons in an industry for abnormal behavior to generate alert. Furthermore, we want to focus on integration of other modules such smoke and fire detection to our framework to propose a complete system working in an industry.

References

[1] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo, "A survey on facilities for experimental internet of things research," *IEEE Communications Magazine*, vol. 49, pp. 58-67, 2011.

[2] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, pp. 2787-2805, 2010/10/28/ 2010.

[3] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, "Future Internet: The Internet of Things Architecture, Possible Applications and Key Challenges," in *2012 10th International Conference on Frontiers of Information Technology*, 2012, pp. 257-260.

[4] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, pp. 1645-1660, 2013/09/01/ 2013.

[5] R. V. Kulkarni, A. Forster, and G. K. Venayagamoorthy, "Computational Intelligence in Wireless Sensor Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 13, pp. 68-96, 2011.

[6] D. E. O'Leary, "BIG DATA', THE 'INTERNET OF THINGS'AND THE 'INTERNET OF SIGNS," *Intelligent Systems in Accounting, Finance and Management*, vol. 20, pp. 53-65, 2013.

[7] L. Da Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Transactions on industrial informatics*, vol. 10, pp. 2233-2243, 2014.

[8] (Accessed on 27 May 2019). Available: https://www.qualitymag.com/articles/94121-machine-visions-central-role-in-the-industrial-internet-of-things

[9] (Accessd on 27 June 2019). Available: https://www.photonics.com/Article.aspx?AID=62511

[10] C. Luo, "Video Summarization for Object Tracking in the Internet of Things," in *2014 Eighth International Conference on Next Generation Mobile Apps, Services and Technologies*, 2014, pp. 288-293.

[11] A. Sehgal, V. Perelman, S. Kuryla, and J. Schonwalder, "Management of resource constrained devices in the internet of things," *IEEE Communications Magazine*, vol. 50, pp. 144-149, 2012.

[12] N. Rahim, J. Ahmad, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Privacy-preserving image retrieval for mobile devices with deep features on the cloud," *Computer Communications*, vol. 127, pp. 75-85, 2018/09/01/ 2018.

[13] B. Kantarci and H. T. Mouftah, "Trustworthy sensing for public safety in cloud-centric internet of things," *IEEE Internet of Things Journal*, vol. 1, pp. 360-368, 2014.

[14] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 678-708, 2015.

[15] M. Thibaud, H. Chi, W. Zhou, and S. Piramuthu, "Internet of Things (IoT) in high-risk Environment, Health and Safety (EHS) industries: A comprehensive review," *Decision Support Systems*, vol. 108, pp. 79-95, 2018.

[16] J. A. Guerrero-ibanez, S. Zeadally, and J. Contreras-Castillo, "Integration challenges of intelligent transportation systems with connected vehicle, cloud computing, and internet of things technologies," *IEEE Wireless Communications*, vol. 22, pp. 122-128, 2015.

[17] T. N. Pham, M. Tsai, D. B. Nguyen, C. Dow, and D. Deng, "A Cloud-Based Smart-Parking System Based on Internet-of-Things Technologies," *IEEE Access*, vol. 3, pp. 1581-1591, 2015.

[18] C. Zhu, C. Zheng, L. Shu, and G. Han, "A survey on coverage and connectivity issues in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 35, pp. 619-632, 2012/03/01/ 2012.

[19] L. Zhang, L. Sun, W. Wang, and Y. Tian, "KaaS: A standard framework proposal on video skimming," *IEEE Internet Computing*, vol. 20, pp. 54-59, 2016.

[20] A. Mahapatra, P. K. Sa, and B. Majhi, "A multi-view video synopsis framework," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 1260-1264.

[21] Y. Muramatsu, T. Hirayama, and K. Mase, "Video generation method based on user's tendency of viewpoint selection for multi-view video contents," in *Proceedings of the 5th Augmented Human International Conference*, 2014, p. 1.

[22] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," *IEEE Transactions on Multimedia*, vol. 12, pp. 717-729, 2010.

[23] S. H. Ou, Y. C. Lu, J. P. Wang, S. Y. Chien, S. D. Lin, M. Y. Yeti, *et al.*, "Communication-efficient multi-view keyframe extraction in distributed video sensors," in *2014 IEEE Visual Communications and Image Processing Conference*, 2014, pp. 13-16.

[24] Y. Li and B. Merialdo, "Multi-video summarization based on Video-MMR," in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, 2010, pp. 1-4.

[25] R. Panda, A. Dasy, and A. K. Roy-Chowdhury, "Video summarization in a multi-view camera network," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, 2016, pp. 2971-2976.

[26] R. Panda, A. Das, and A. K. Roy-Chowdhury, "Embedded sparse coding for summarizing multi-view videos," in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2016, pp. 191-195.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.

[28] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. d. Albuquerque, "Cloud-Assisted Multi-View Video Summarization using CNN and Bi-

Directional LSTM," *IEEE Transactions on Industrial Informatics*, pp. 1-1, 2019.
[29] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66-72, 2018/01/31/ 2018.
[30] J. L. Lobo, J. Del Ser, I. Laña, M. N. Bilbao, and N. Kasabov, "Drift Detection over Non-stationary Data Streams Using Evolving Spiking Neural Networks," in *International Symposium on Intelligent and Distributed Computing*, 2018, pp. 82-94.
[31] A. Ullah, K. Muhammad, J. D. Ser, S. W. Baik, and V. H. C. d. Albuquerque, "Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM," *IEEE Transactions on Industrial Electronics*, vol. 66, pp. 9692-9702, 2019.
[32] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," *IEEE Access*, vol. 6, pp. 1155-1166, 2018.
[33] M. Sajjad, S. Khan, T. Hussain, K. Muhammad, A. K. Sangaiah, A. Castiglione, *et al.*, "CNN-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognition Letters*, 2018.
[34] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91-99.
[36] ((Accessed on 28 June 2019)). Available: https://docs.opencv.org/master/d4/da8/group__imgcodecs.html#gga292d81be8d76901bff7988d18d2b42acad2548321c69ab9c0582fd51e75ace1d0
[37] G. Wallace, "The JPEG Still Picture Compression Standard Communication of the ACM 34," 1991.
[38] S.-H. Ou, Y.-C. Lu, J.-P. Wang, S.-Y. Chien, S.-D. Lin, M.-Y. Yeti, *et al.*, "Communication-efficient multi-view keyframe extraction in distributed video sensors," in *Visual Communications and Image Processing Conference, 2014 IEEE*, 2014, pp. 13-16.
[39] S. K. Kuanar, K. B. Ranga, and A. S. Chowdhury, "Multi-view video summarization using bipartite matching constrained optimum-path forest clustering," *IEEE Transactions on Multimedia*, vol. 17, pp. 1166-1173, 2015.
[40] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, pp. 165-179, 2015.
[41] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *arXiv preprint arXiv:1706.03121*, 2017.

Tanveer Hussain (S'19) received his Bachelor's degree in Computer Science from Islamia College Peshawar, Peshawar, Pakistan in 2017. He is currently pursuing his M.S. leading to Ph.D. degree from Sejong University, Seoul, Republic of Korea and serving as Research Assistant at Intelligent Media Laboratory (IM Lab). His major research domains are features extraction (learned and low-level features), video analytics, image processing, pattern recognition, deep learning for multimedia data understanding, single/multi-view video summarization, IoT, IIoT, and resource constrained programming.

Khan Muhammad (S'16–M'18) received the Ph. D degree in Digital Contents from Sejong University, South Korea. He is currently an Assistant Professor in the Department of Software, Sejong University, South Korea. His research interests include medical image analysis (brain MRI, diagnostic hysteroscopy and wireless capsule endoscopy), information security (steganography, encryption, watermarking and image hashing), video summarization, computer vision, fire/smoke scene analysis, and video surveillance. He has published over 60 papers in peer reviewed international journals and conferences in these research areas with target venues as IEEE COMMAG, Networks, TII, TIE, TSMC-Systems, IoTJ, Access, TSC, Elsevier INS, Neurocomputing, PRL, FGCS, COMCOM, COMIND, JPDC, PMC, BSPC, CAEE, Springer NCAA, MTAP, JOMS, and RTIP, etc. He is also serving as a professional reviewer for over 40 well-reputed journals and conferences.

Javier Del Ser (M'07-SM'12) received his first PhD in Telecommunication Engineering (Cum Laude) from the University of Navarra, Spain, in 2006, and a second PhD in Computational Intelligence (Summa Cum Laude) from the University of Alcala, Spain, in 2013. He is currently a Research Professor in data analytics and optimization at TECNALIA (Spain), a visiting fellow at the Basque Center for Applied Mathematics (Spain) and an adjunct professor at the University of the Basque Country UPV/EHU. His research activity gravitates on the use of descriptive, prescriptive and predictive data mining and optimization in a diverse range of application and sectors such as Energy, Transport, Telecommunications, Industry and Security, among others. In these fields he has published more than 220 articles and

conference contributions, co-supervises more than 10 Ph.D. theses, has edited 4 books and co-invented 6 patents.

Sung Wook Baik (M'16) received the B.S degree in computer science from Seoul National University, Seoul, Korea, in 1987, the M.S. degree in computer science from Northern Illinois University, Dekalb, in 1992, and the Ph.D. degree in information technology engineering from George Mason University, Far fax, VA, in 1999. He worked at Datamat Systems Research Inc. as a senior scientist of the Intelligent Systems Group from 1997 to 2002. In 2002, he joined the faculty of the College of Electronics and Information Engineering, Sejong University, Seoul, Korea, where he is currently a Full Professor and the Chief of Sejong Industry-Academy Cooperation Foundation. He is also the head of Intelligent Media Laboratory (IM Lab) at Sejong University. His research interests include computer vision, multimedia, pattern recognition, machine learning, data mining, virtual reality, and computer games.

Victor Hugo C. de Albuquerque (M'17–SM'19) received the graduation degree in mechatronics technology from the Federal Center of Technological Education of Ceará, Fortaleza, Brazil, in 2006, the M.Sc. degree in tele-informatics engineering from the Federal University of Ceará, Fortaleza, in 2007, and the Ph.D. degree in mechanical engineering with emphasis on materials from the Federal University of Paraíba, João Pessoa, Brazil, in 2010. He is currently an Assistant VI Professor with the Graduate Program in Applied Informatics at the University of Fortaleza, Fortaleza. He has experience in computer systems, mainly in the research fields of applied computing, intelligent systems, visualization and interaction, with specific interest in pattern recognition, artificial intelligence, image processing and analysis, Internet of Things, Internet of Health Things, as well as automation with respect to biological signal/image processing, image segmentation, biomedical circuits, and human/brain–machine interaction, including augmented and virtual reality simulation modeling for animals and humans.